Carnegie Mellon University

TESTIMONY BEFORE THE
SUBCOMMITTEE on CONSUMER PROTECTION, PRODUCT SAFETY and DATA SECURITY
SENATE COMMITTEE on COMMERCE, SCIENCE and TRANSPORTATION
HEARING ENTITLED, "THE NEED FOR TRANSPARENCY IN ARTIFICIAL INTELLIGENCE"

Ramayya Krishnan
W. W. Cooper and Ruth F. Cooper Professor of Management Science and Information Systems
Dean, Heinz College of Information Systems and Public Policy
Founding Faculty DIrector, The Block Center for Technology and Society
Carnegie Mellon University

Sep 12, 2023

Chair Cantwell, Ranking Member Cruz, Subcommittee Chair Hickenlooper, Ranking Member Blackburn and Members of the Committee, I am grateful for the opportunity to testify today. My name is Ramayya Krishnan and I serve as Dean of the Heinz College of Information Systems and Public Policy, a multidisciplinary academic unit that spans information technology, data analytics and public policy at Carnegie Mellon University. My perspective is shaped by my work over several decades on the use of data analytics and advanced model-based approaches to support decision making in consequential applications and by my role as the faculty director of the Block Center for Technology and Society, a university wide initiative that studies the responsible use of AI and its consequences for the future of work. I am a member of the National AI Advisory Committee (NAIAC). However, I am here in my own capacity and not representing the NAIAC.

Artificial intelligence (AI) is a transformative technology. Its application to create personalized tutors (See https://youtu.be/yEgHrxvLsz0), create operational and clinical decision support tools in health care (https://www.nature.com/articles/s41586-023-06160-y), promote health literacy among citizens https://www.cmu.edu/news/stories/archives/2023/august/revolutionizing-health-care-harnessing-artificial-intelligence-for-better-patient-care), and enable breakthroughs in science and drug discovery that will unlock solutions to currently intractable problems in human health and beyond are among the many economically and societally beneficial uses of the technology. And likely many of us have used an AI chatbot like chatGPT, a generative AI technology, and seen both its immense potential and its failures.

As AI technologies are considered for use in high stakes applications such as in health care, recruiting and criminal justice, the unwillingness of the leading vendors to disclose the attributes and provenance of the data they have used to train and tune their models, and the processes they have employed for model training and “alignment” to minimize the risk of toxic or harmful

responses needs to be urgently addressed (https://arxiv.org/pdf/1901.10002.pdf). This lack of transparency creates threats to privacy, security, and uncompensated use of intellectual property and copyrighted content in addition to harms caused to individuals and communities, due to biased and unreliable performance. There is a need for greater accountability and transparency in the development and deployment of AI to spur its responsible adoption and use.

In my testimony, I propose four decisive recommendations for Congress to consider to address these challenges. Investing in these recommendations will provide near term impact on trusted adoption of AI and, when combined with a focused research program, will ensure US leadership in responsible and trustworthy AI.

The recommendations include foundational actions to advance broad based adoption of Responsible AI practices as well as measures to accelerate the adoption of practices and technologies to mitigate the threat of deep fakes and protect the rights of creators. My final recommendation is to create an infrastructure for AI trust by establishing a capability to monitor and respond to AI vulnerabilities and failures and support the development and dissemination of solutions and best practices. These measures will need to be closely aligned with a focused research program that advances the development of tools for watermarking and labeling as well as research into measurement, metrics and evaluation of reliability and quality of the AI supply chain. A detailed policy memo co-authored by my colleagues and myself from Carnegie Mellon on accountable AI is available at https://www.cmu.edu/block-center/responsible-ai/cmu_blockcenter_rai-memo_final.pdf.

1. **Promoting Responsible AI**

   Congress should require all federal agencies to use the NIST (National Institute of Standards and Technology) AI Risk Management Framework during the design, development, procurement, use, and management of their AI use cases. This will promote responsible adoption and deployment of AI in government and more broadly in society. Investing in workshops such as the NIST- Carnegie Mellon AI RMF workshop which convened academics and industry representatives from sectors such as banking, health care and consulting to discuss the gaps that need to be addressed to operationalize responsible AI in the economy will be particularly valuable. The NIST AI RMF was developed with multiple stakeholder inputs and establishing it as a standard will have numerous benefits at home and abroad.

2. **Promoting Greater AI Transparency**

a. **Content transparency: Content Labeling and Detection:**

   A significant concern with the advent of generative AI is the ease with which multi-modal content (audio, video text) can be created that is indistinguishable from human created

content. Multiple examples document why this is a problem. Students submitting AI-produced content in lieu of their own work is an academic integrity issue and hurts learning outcomes. Audio and video deep fakes raise concerns from multiple standpoints - from affecting the economic outcomes and reputations of well-known artists to concerns for human rights and democracy.

Currently, there is no standardized way to label content as AI generated and no standardized tool that can use such labels to help consumers recognize AI generated content. Recent commitments by major vendors to develop watermarks for AI generated content and the emergence of content provenance standards (e.g., C2PA) is a step in the right direction. While proposals exist for audio and visual content, watermarking and provenance for AI generated text remains a challenge. As with all security technologies, watermarking will need to stay one step ahead of attempts to defeat it (https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/ and https://gjia.georgetown.edu/2023/05/24/should-the-united-states-or-the-european-union-follow-chinas-lead-and-require-watermarks-for-generative-ai/).

While the usual concern about this issue has been from the point of view of the consumer, this inability to distinguish AI generated content from human produced content and knowledge of its provenance is relevant to model developers as well. Since internet content is used at scale to train models and as AI produced content proliferates on the internet, model developers will need the capacity to differentiate AI produced content from human produced content since this has implications for model performance.

Congress should require all AI models (open source and closed source models) that produce content to label their content with watermarking and provenance technology and provide a tool to detect the label.

b. Advancing Transparency in the AI pipeline for high stakes applications

The AI pipeline or value chain (Hosanagar and Krishnan, 2023) consists of training data, models and applications. It is this pipeline that is used to create AI systems in high stakes applications such as in autonomous vehicles, health care, recruiting, and criminal justice. The leading vendors of closed source models do not disclose the attributes and provenance of the data they have used to train and tune their models, and the processes they have employed for model training and “alignment” to minimize the risk of toxic or harmful responses. When these AI pipelines are used in high stakes applications, greater transparency around how AI is integrated into the broader system is needed. We can learn from prior work in accountable AI as well as in modeling reliability of societal and engineered systems to address these transparency questions. In the following, I will highlight key recommendations that pertain to data transparency and AI model validation and evaluation. These measures are vital to address critical challenges created by Large

Language Models. It will also be vital in enabling U.S. companies to remain globally competitive as international standards are developed.

- **Advancing Data Transparency**

- Model developers need to document the rights they have to work with the data they are using to train the model. This documentation should also provide information about the source of the data, whether it was public or private, etc.
- Model developers should respect the right of data owners to opt out of data crawling (robots.txt file) and also provide data owners the opportunity to opt out of the use of their already collected data in model training or tuning.
- Model developers need to document the standards that were used in bias assessment and demonstrate the analysis that was conducted to assess structural bias in the data.

Congress should require standardized documentation and, like audited financial statements, they should be verifiable by a trusted third party (e.g., an auditor). The metaphor is to think of these as akin to "nutrition labels." so it is clear what went into producing the model.

- **Promoting Model Validation and Evaluation of the AI system**

- Develop clear standards for articulating intended use cases and metrics for reliability and utility so that users can have clear expectations of performance under well-defined conditions. **Congress should direct NIST to develop standards for these important societal domains.**
- Define AI sandboxes and test data sets, evaluation frameworks, measurement and metrics, and continuous monitoring standards based on the assessed risk of the application space or use case.
- Require the auditor to use these standards and validation infrastructure to evaluate the AI system and provide the required assurance prior to deployment.

Congress should require a model validation report for AI systems deployed in high stakes applications. The metaphor is to think of this as akin to an "underwriters lab" that objectively assesses the risk and performance of an AI system.

c. Investing in a trust infrastructure for AI (AI CERT), or ALRT (AI lead response team)

AI is developing rapidly. Even with the proposed transparency measures, there will be a need to respond rapidly to newly discovered AI vulnerabilities, exploits and failures. The safety and reliability of the AI ecosystem is a necessary condition to engender trust and spur widespread adoption and deployment of AI. While AI Incident databases (https://incidentdatabase.ai/) from the Responsible AI Collaborative, Project Atlas from

MITRE (see https://atlas.mitre.org/) and the recently organized DEFCON red teaming event and voluntary commitments are important steps forward, an institutional solution is required.

The proposed solution, an ALRT, would connect vendors, AI system deployers and users. It would catalog incidents, record vulnerabilities, test and verify models, and recommend solutions and share best practices to minimize systemic risks (https://www.forbes.com/sites/rosecelestin/2023/08/23/the-ai-financial-crisis-theory-demystified-how-to-create-resilient-global-ecosystems/?sh=27282b4d51ce) as well as harm stemming from vulnerability exploits. This is modeled after the computer emergency response team (CERT) that the US Government stood up in response to cyber security vulnerabilities and threats in late 1980s. The following capabilities are required to serve the needs of CERT for AI)

- Catalog incidents, record vulnerabilities, recommend solutions and, share best practices to minimize risks
- Coordinate commercial and public government focus with the need to rapidly respond to national security challenges (e.g., chemical/bio weapons https://arxiv.org/ftp/arxiv/papers/2306/2306.03809.pdf). Be able to respond to threats that affect .com, .gov and .mil domains. In effect, combine open, restricted and classified work
- Possess deep technical capabilities spanning core AI and computing and an understanding of how the core AI is operationalized to meet application needs
- Maintain domain knowledge connected to applications
- Convene industry, government and academic partners around core tech AI technology as well as operationalization of AI

Congress should stand up these capabilities quickly via existing FFRDCs and harness the strengths at NIST and other federal agencies. implementation of these recommendations will have an immediate impact on trusted adoption of AI (e.g., standing up ALRT). Combining implementation with investments in a focused program of research on Responsible AI and AI transparency will ensure US leadership in trustworthy AI.

Finally, in closing, the success of these recommendations will in part rest on comprehensive strategies that enhance AI skills across the continuum from K-12 and community college education to new tools, strategies and policies to support workers in virtually all industries to adapt to the impact of AI. Thank you for this opportunity to testify to your committee.